Education Policy Analysis Archives

Volume 10 Number 10 January 28, 2002 ISSN 1068-2341

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Response to Michelson and to Willson and Kellow

Craig Bolon
Planwright Systems Corporation (USA)

Citation: Bolon, C. (2002, January 28). Response to Michelson and to Willson and Kellow. *Education Policy Analysis Archives*, *10*(10). Retrieved [date] from http://epaa.asu.edu/epaa/v10n10/.

Abstract
The criticisms and points made by both Michelson (http://epaa.asu.edu/epaa/v10n8/) and Willson and Kellow (http://epaa.asu.edu/epaa/v10n9/) in response to my article "Significance of Test-based Ratings" (http://epaa.asu.edu/epaa/v9n42/) are here addressed.

Michelson's Complaints

Michelson's critique of "Significance of Test-based Ratings" rides herd on some fine points but misses main themes of the article. The article's data include test scores for only 47 schools. As experts have warned, such a small behavioral data set can typically provide stable coefficients for only one or two independent variables. The work leading to the article aimed to see if one or two strong variables could be found for this limited

data set. As it happened, a dominant variable was found: community income.

In seeking an expanded analysis, Michelson overloads the observations with independent variables, adding some with no evidence for underlying quality. While his approach associates more variance with a larger set of variables, he does not provide stepwise or combinatorial analysis for the incremental association of variance or conduct a sensitivity study to explore the likelihood that his results may be an artifact. Despite an undiscriminating approach, community income remains the strongest factor.

The article, by contrast, emphasizes robust results obtained from accurate, traceable data and parsimonious models. It employs cross-validation, sensitivity analysis and combinatorial analysis. Weighting is introduced to construct models that will not be destabilized by smaller schools contributing to the data set.

Apparently unsatisfied with his alternative, since community income is still the strongest factor, Michelson proceeds with polemics centered around the notion that the article really has no news anyway, since (somehow) everybody knows that high test scores go along with high incomes. Maybe everyone in his circle does, but many people I encounter are surprised; they wonder why this should be so.

It is known, if not well known, that by the early 1920s labor unions mounted protests against the social injustice of using IQ scores to place students in academic "tracks." They had found out that high IQ scores were strongly associated with high family incomes. It is also known, if not well known, that by the mid-1950s the Educational Testing Service had found a regular progression of their average SAT scores with average reported family incomes. But these results are for "aptitude" tests.

There have been limited published studies about the associations of social and school factors with state "accountability" test scores. Will such a test be similar in social correlates to "aptitude" tests, or will it be different? Massachusetts is a useful laboratory for such a study. It did not begin "accountability" testing until 1998. It then created what is generally regarded as a state-of-the-art program.

The Massachusetts graduation requirement will not directly affect any student until 2003. Although Massachusetts has some communities with a history of aggressive testing, such as Worcester, before the school year ending in 2000 most schools made light to moderate responses to MCAS tests. These circumstances help provide a good baseline.

Problems for such studies are the rare availability of reliable, personal social-factors data and the limited social-factors data that one can clearly associate with individual schools. Most data collected and reported by schools either count disadvantaged populations or count eligibility for free or reduced-price lunch. This produces two common outcomes. One is correlations found for test scores with population categories. The other is correlations found between test scores and poverty, since poverty or near-poverty income is the qualification for free or reduced-price lunch.

Beginning with Massachusetts school profiles, I also found those correlations. But being familiar with the communities for which I had data, I decided to look at residuals. To me the residuals seemed to show a pattern—high score-residuals in high-income communities and vice-versa. Data for disadvantaged populations, poverty and other

school categories did not appear to tell the whole story, so I sought income data for school populations.

It turned out that the only generally available data were from the 1989 U.S. Census of Population and Housing. Somewhat to my surprise, per-capita community income from a decade prior to the test scores proved to be a strong and robust factor. The article recounts the modeling of data in the sequence it occurred.

The major theme, which Michelson does seem to understand, is that income appears to matter at all ordinary levels, not just at the threshold of poverty. Part of this may be self-fulfilling prophecy, when test scores are used to grant or deny advancement, but there is probably more. I don't accept Michelson's guesses about the phenomenon. I have different hypotheses but won't trust those either without evidence.

Another theme, which Michelson seems to ignore, is that community income, as distinct from family income, may have a powerful effect. That is merely a suggestion in the article, of course. It would take a study of individuals to differentiate the influences.

A fortuitous circumstance for this study was the pattern of New England cities and towns, which form legal boundaries around small, diverse clusters of population. That can also be found elsewhere in the U.S., such as near Philadelphia or Cleveland, but in the more recently settled areas it is rare. Instead, a large city has usually been allowed to swallow up many neighbors, and the remaining suburbs do not have the diversity of the urban neighborhoods. Social data collected within city boundaries can be very difficult to reaggregate, as happens in the City of Boston.

The following paragraphs respond to Michelson's observations item by item:

(1) Michelson first complains about what he calls a "non-sequitur." The article's abstract says, "The state [of Massachusetts] is treating scores and ratings as though they were precise educational measures of high significance." And indeed it is doing just that. As the article later points out, getting 23 instead of 24 correct answers on a tenth-grade mathematics test can be enough for Massachusetts to deny a student a high-school diploma. Such a small difference can produce a huge effect, since high-school graduation has great influence on lifetime income.

The article's abstract goes on to say that a "review of tenth-grade mathematics test scores...showed that statistically [Massachusetts scores and ratings] are not [precise educational measures of high significance]." And indeed the scores are not precise. As the article shows, the variability of score averages is so large that at least several years will be needed to see whether there are definite trends for most schools. As the article also shows, the educational significance of the scores is highly questionable. As with "aptitude" tests, scores closely track income levels. Once predictions based on income have been subtracted, few schools can be distinguished. There is little to indicate that these scores may measure what schools achieve, as contrasted with what social advantages or disadvantages students bring to schools from their backgrounds.

(2) Michelson complains the article "has not said what success on these tests is supposed to imply." That's not my job; it's the job of an agency in charge of the tests. What the article says is that no studies "have shown that MCAS test scores have practical significance, in the sense of predicting success in adult activities to any greater degree

than could be done with knowledge of student backgrounds."

It should be a responsibility of government, when using tests so as to cause drastic life consequences for young people under its care, to demonstrate objectively that its tests accurately and fairly measure skills of critical and lasting importance. Massachusetts has failed to do this. It merely says its tests are similar to other tests, which also lack practical validation. A physician acting in such a cavalier way toward a patient would be at risk of fines or jail.

The topic of the article, however, is school ratings. We maintain public schools to equip all young people with skills and knowledge essential to support themselves and to carry out civic responsibility. Schools in rich and poor communities alike strive to do this. School ratings that mostly track incomes of communities are unlikely to reflect actual levels of effort or achievement by the schools. The Massachusetts tests lack practical validation, and the ratings based on those tests appear to measure characteristics of communities more than they do those of schools.

(3) Michelson complains about dropping Boston schools from part of the data analysis. That part of the analysis focuses on community income. As the article indicates, given Boston's complex mix of exam schools, magnet schools, district schools, cross-enrollment and busing, it was not possible to determine community income for individual schools. If one wants to expand the data set, it would more fruitful to add other Massachusetts cities and towns than to spend the large amount of effort needed to reaggregate Boston data by schools with any accuracy.

(4) Michelson complains about weighting by number of test takers, but aside from an appeal to prejudice he does not try to explain why an unweighted analysis would be of more use. For example, one way to handle complaint (3) might be to aggregate all of Boston's schools and use citywide per-capita income and other factors. Would it then be useful to treat all of Boston, with a population of 589,141, as equivalent in weight to Winthrop, with a population of 18,303? Despite this grievance, as Michelson later shows, unweighted analysis leads to similar patterns of results. No reviewer of this article raised concerns about weighted models.

(5) Michelson objects to lack of a marker variable for schools with vocational programs in the same facility as academic programs. However, some such schools have only vestigial programs, while other schools enroll large fractions of their students in vocational programs, with year-to-year changes depending on local circumstances. Massachusetts school profiles did not record these programs uniformly when study data were assembled and do not provide program enrollments. A marker variable can be wildly inaccurate. Later Michelson also proposes a marker variable for exam schools. Boston's exam schools are known to vary widely in selectivity, and a marker variable will not account for that.

(6) Michelson's complaint about weighting by numbers of test takers while also including school population as a variable in one of the analysis steps is reasonable; the correlation is quite high. However, I supplied Michelson with a file of all data used for the article. As I saw and he should readily have found, both weighted and unweighted analysis show low significance for this variable. I did not think that putting an additional analysis into the article would add much information, but perhaps a comment about checking unweighted analysis should have appeared.

(7) Michelson's complaint about using only "per pupil expenditure" (regular education) as an estimator for financial support is directed to the wrong party. Massachusetts citizens have protested for years about poor reporting of school spending. As the article says in an appendix, Massachusetts is finally adopting a uniform system of detailed financial reporting. The first published data from the new system will be available some time in 2004.

(8) Michelson makes an attempt to estimate cross-effects of Boston's exam school system in lowering scores of district schools. However, as the article says, there is evidence that many ambitious parents whose offspring who are not accepted to an exam school of their choice send them outside Boston public schools: to parochial schools, other private schools and suburban schools. It would take far more resources than the data for this study provide to investigate cross-enrollment effects accurately.

(9) Michelson objects to the use of unadjusted R^2 to report the variance associated in successive steps of analysis. This would be a reasonable complaint if there were no dominant variable and one needed to account in detail for relative contributions of multiple variables, but that was not the result found in the study. Using adjusted R^2 would intensify the dominance of the community income variable, because adjusted R^2 values obtained when adding more variables to a model are lower than unadjusted values. In his use of adjusted R^2, Michelson does not show that the assumptions of adjustment are actually satisfied for the data.

(10) Under "A Change of Method," Michelson again objects to dropping Boston schools from part of the analysis and seems to miss the point where the change occurs. He states that text following Table 2-10 and a figure in Table 2-13 report an unadjusted R^2 value of .80 for three variables, while Table 2-15, he says, reports a value of .86 for only two variables. This seems to puzzle him; he says he can "see no reason for" it.

Actually, the R^2 value of .86 for two variables is reported in the paragraph preceding Table 2-14. This paragraph begins by saying that analysis starting with that table applies "only to schools outside the City of Boston." A possible cause of the increase in R^2 values is, as the article says, that one cannot accurately determine community income for individual Boston schools. When analyzing mainly schools with unambiguous community income, one is looking at a less noisy data set. The procedure and the reason for it are clearly stated.

(11) In his objections to the article's use of unadjusted R^2, Michelson makes much of the difference in R^2 values between the analyses that include the Boston schools and those that don't. Unlike Michelson, the article does not try to compare R^2 between these analyses. They don't use the same data set. Without knowing what is not known about the social factors for individual Boston schools, comparisons like those Michelson suggests will not be meaningful.

(12) Michelson also seems unsure when weighted or unweighted analysis is being used. Weighting by number of test participants is described in the paragraph preceding Table 2-3, the first analysis reported in the article, and was used for that and all following

analyses (except, as the article states, those in Figure 2-3). Readers are occasionally reminded that models are being weighted in this way. If Michelson uses unweighted analysis or a different weighting factor, such as school population, then he will get different results.

(13) Michelson's Table 2 and Table 3 offer what he calls an "all schools replication" of analyses in Table 2-14 and Table 2-16 of the article. The two paragraphs preceding Table 2-14 in the article discuss problems of reaggregating data for Boston schools and introduce analyses that consider only the remaining schools, reported in Tables 2-14 through 2-16. An "all schools replication" is not a replication. Since he doesn't use the same data set in his Tables 2 and 3, Michelson gets different results. When he uses the same data set in his Tables 4 and 5, he gets the same results as the article contains.

Michelson notes a concern about excluding only Boston schools in Table 2-14 and Table 2-16, since Quincy, Lynn and Newton also have multiple high schools. The article makes the same observation in its summary analysis, Section 2.D, and presents in Tables 2-20 and 2-21 and in Figures 2-8 through 2-10 results excluding all four of these communities from the data set.

(14) Again, Michelson objects to dropping Boston schools from the data set and proposes a different model into which he introduces marker variables, "Michelson's 5-Factor Model." I considered such an approach during the study but rejected it for the reasons stated under his complaint (5): these marker variables may wildly misrepresent what they claim to identify. In rejecting weighting for his model, Michelson exposes it to instability from smaller schools that are well off the trend lines. He does not explain his reasons for the choice.

Here, from a reader's perspective, Michelson is exploring new ground. He has different techniques for analyzing different sets of data and seems to have different motives. His model ignores conservative recommendations for behavioral data by using too many variables in a final model for the size of the data set. He does not try to overcome the potential problems in this approach with a sensitivity study to explore probable ranges of results from such a model. He does not review any potential weakness of his marker variables. He does not provide readers with stepwise or combinatorial analysis for incremental association of variance, only first and last steps. He does not attempt any cross-validation. He does not explain his rejection of weighting. The results shown in Michelson's Figure 1 may represent a robust pattern, or they may be a statistical artifact.

As previously stated, the article emphasizes robust results obtained from accurate, traceable data and parsimonious models. It employs cross-validation in Table 2-13, sensitivity analysis in the exploration of outliers in and near Tables 2-18 and 2-19, and combinatorial analysis in Table 2-12. Weighting is introduced beginning at Table 2-3 to construct models that will not be destabilized by smaller schools contributing to the data set.

(15) Besides offering no stepwise or combinatorial analysis of his own, Michelson objects to such analysis in the article, shown in Table 2-12 and 2-15 and discussed in Section 2.C. A curious objection, since stepwise and combinatorial analyses are common, helpful approaches to understanding the relative influences of multiple factors. A better objection would have been to call at this point for the use of adjusted R^2, since when the assumptions of adjustment can be satisfied, the adjustment will discount added

factors of low significance.

(16) Michelson characterizes the article's discussion of limited English proficiency as "political" and claims the article largely ignores it. Readers can judge for themselves. Section 2.C of the article says, "The factor ‘Percent limited English proficiency' was the second strongest influence on predicted test scores." It offers hypotheses for further investigation suggested by this finding. It then goes on to discuss relative significance of factors "Percent African American," "Percent Hispanic / Latino," and "Percent Asian or Pacific Islander." (Unlike Michelson, I prefer longer, more informative factor names to "bosnoex" and other mysterious abbreviations.)

(17) Michelson claims the article ignores what he calls "specification" effects in the evolution of model equations. In fact, several parts of the article address just such effects, and the article as a whole is an extended model development.

In particular, the discussion of Table 2-6 emphasizes how an economic factor captures variance otherwise associated with disadvantaged populations. Table 2-7 is presented to show how a model without the economic factor loads significance onto population categories. The analysis in Table 2-12 shows how different model equations reveal the weakness of one factor. Discussion around Table 2-14 shows how the factor "Per-capita community income (1989)" supplants the significance of the factor "Percent free or reduced price lunch." (Unlike Michelson, I assume readers are familiar with such effects and do not need a lecture.)

(18) Arguing about his marker variable for vocational programs, Michelson ignores the lack of data characterizing the programs or the students who enroll in them. It might be that the programs draw many students from low-income families or families who do not speak standard English as a first language. It might be that the programs neglect skills or knowledge being tested by MCAS. There could also be a combination of these factors, or there might be some critical but wholly different factor. Data available for the study were insufficient to address the issue, and the article says so. I believe it is unwise for Michelson to introduce a marker variable without investigating the environment.

(19) Similarly, arguing about his marker variable for Boston exam schools, Michelson relies on personal recollections and anecdotes, but he ignores the complex social characteristics of Boston and the lack of reliable data for estimating cross-enrollment effects, which only begin with the exam schools. A study by the Mumford Center at the University of Albany indicates that parochial and other private schools have such large effects that the incidence of poverty among the households of Boston public school students substantially exceeds the incidence in the city population. As with vocational programs, data available for the study were insufficient, and the article says so. Again, I believe it is unwise for Michelson to introduce a marker variable without investigating the environment.

(20) Caught up in personal anecdotes, Michelson ignores findings about individual communities reported in the article that appear significant. Belmont substantially outscored predictions, while Marblehead scores were considerably lower than predicted. Sensitivity analysis suggests robust results, not artifacts. Neither community is known for extremes in population or education; a review that compares and contrasts them might be of interest and of practical use.

(21) Michelson's objections to the residuals discussion around Figures 2-4 and 2-5 in the article ignores the increase in slope of the line of fit. These figures were placed in sequence so that this effect could be easily seen. Michelson is correct in his observation that annual score averages and score changes are limited predictors—the point made with Figures 2-9 and 2-10 in the article. With Michelson's scatterplot of successive year residuals from "Michelson's 5-Factor Model" in his Figure 3, he might raise a question about the anomalous behavior of Swampscott, which the article called attention to in discussing Table 2-11.

(22) Under "The Town View," Michelson begs the question of the article. He contends that "the highest scoring students are exactly who we would expect them to be, those from the highest income places...." Yes, if you look at the plot of average scores versus income in the abstract of the article, that is what you would expect. But if you hadn't seen the data, would you know? Perhaps, as he seems to suggest, Michelson is privy to inside information. Most of us have to look at the data to find out.

(23) Michelson's "Final Remarks" include the polemics previously mentioned. In these, he contends that "MCAS tests are designed to measure individual achievement" and seems to want to make this an affair of honor. True or false, meaningful or otherwise, that's not the topic of the article, which also seems to have escaped Michelson. Likewise, the article is concerned neither with what Michelson calls "beliefs" about its findings nor (certainly, in an article using statistical inference) with causality.

Summary of Response to Michelson

The topic of the article, "Significance of Test-based Ratings for Metropolitan Boston Schools," is the meaning and usefulness of school ratings that are entirely based on MCAS test scores. The article shows, for the years and tests it reports, that within the variations of test score averages the Massachusetts Department of Education could have produced nearly the same ratings simply by scaling income data from the Department of Revenue. As an appendix to the article notes, Section 1I in Chapter 69 of the Massachusetts General Laws directs the Department of Education to set up a school rating system with a broader approach than it has used so far:

> "The system shall employ a variety of assessment instruments on either a comprehensive or statistically valid sampling basis. Such instruments...shall include consideration of work samples, projects and portfolios, and shall facilitate authentic and direct gauges of student performance."

This provision was written when the state was administering tests on a sampling basis that were inspired by NAEP, which tries to acknowledge a variety of learning orientations. In narrowing its current approach to a single test series, Massachusetts may have emphasized only the cognitive skills sampled by "aptitude" tests. Certainly it fails to honor the spirit of its laws.

Unlike the impression Michelson gives of his outlook, I don't see statistical analysis as a card game, playing to get a high multiple R score while discounting the quality of data. Statistics won't identify causes or distinguish causes from effects. At best one can find robust patterns that justify investigation by other means. Knowing, for example, that community income provides a strong, persistent factor for certain test scores may

motivate someone to find out why this happens and might eventually lead to better understanding of how or what to teach or test.

Knowing that factors may have some influence will not help an investigator unless their influence is major. In this vein, someone with a scientific or engineering background will tend to apply a $p<0.05$ criterion as a rough, first cut—the criterion to which Michelson takes such pained exception. That isn't a signal of existential meaning; it's a value judgement. If there isn't much more than a 95 percent chance of significance, then another phenomenon is probably a better object of one's attention. Anyone who tries to chase down all the "Michelson 5-Factor[s]" with surveys or experiments is risking a waste of energy in blind alleys.

Michelson's reported confusions with the article suggest that he would really like to analyze different data sets with different techniques. That's fair, of course, and it might yield some new information. But it really proposes a different article, which Michelson seems to have begun in the guise of a critique.

As "Significance of Test-based Ratings" shows, Massachusetts school ratings, based solely on MCAS test scores, are not precise educational measures of high significance. Score variations are large, and scores appear to reflect heavily the social advantages or disadvantages that students bring to schools from their backgrounds, not necessarily the effectiveness of the schools themselves.

Willson's and Kellow's Concerns

Willson's and Kellow's concerns about "Significance of Test-based Ratings" focus largely on what they say are "theoretical" and standards issues. The article presents data, models and correlations associating state "accountability" test scores used to construct school ratings with social and school characteristics. It does not endorse any "theoretical" framework or invoke any institutional standard of judgement. Willson and Kellow also raise issues about construction and content of tests and discuss data developed by identifying individual students. While these are reasonable subjects of inquiry, they are not the topic of the article.

The article explores the potential significance of school ratings based on "accountability" test scores. The study on which the article was based looked at average test scores and social and school factors for 47 geographically clustered schools to see if one or two strong factors could be found for this limited data set. A dominant factor was found: community incomes. Results show that the test scores track community incomes so closely that it is questionable whether the scores measure efforts or effectiveness of the schools.

Willson and Kellow seem to feel that income inequities should no longer be considered particularly relevant to educational issues, since funding has been equalized. Although both their home state of Texas and mine of Massachusetts have school funding equalization programs, both states also continue to encounter strong disagreements and lawsuits over the issue, for example:

> "In a lawsuit filed against the state of Texas, lower-wealth school districts allege that the state's current funding scheme for education fails to meet the equity and efficiency standards established by a 1994 Texas Supreme Court

> decision. Plaintiffs in the suit claim that tax exemptions for, among other things, country clubs and sports franchises represent lost property tax revenues that would otherwise be allocated to funding the state's schools. These breaks reduce school funding by an estimated $500 million, according to the plaintiffs. A statewide property tax to correct the state's school funding inequities failed in the last session of the Texas legislature." (from Texas Education Funding, 1998)
>
> "Residents Advocating Government Equity, or RAGE, sought $50,000 from Barnstable County to continue a private lawsuit, brought in the name of eight Cape schoolchildren to the Massachusetts Supreme Judicial Court. The suit asks the court to intervene and make the state legislature carry out its constitutional duty to adequately finance aid to schools. That includes changing the aid formula to better serve towns like those on the Cape, with relatively high property values but also with high growth rates. Several Cape towns are also owed $14.7 million in back aid." (from Milton, 1999)

Willson and Kellow don't consider recent evidence such as Grissmer, et al., 2000, that income levels may influence test scores independently of school funding. In Figures 2-6 and 2-7 the article "Significance of Test-based Ratings" shows, for the communities studied, that per-capita community income is strongly associated with average test scores while school spending shows very little association with average test scores.

Willson and Kellow make a vague statement that "all sorts of predictors" (other than income) might produce strong correlations with test scores, but they don't offer any evidence. Why not? If strong predictors were so easy to come across, surely they could dredge up a few. Actually, an appendix to the article provided Willson and Kellow with real data to test such a proposition, the social and school factors that the article analyzes. The results are in Figure 1.

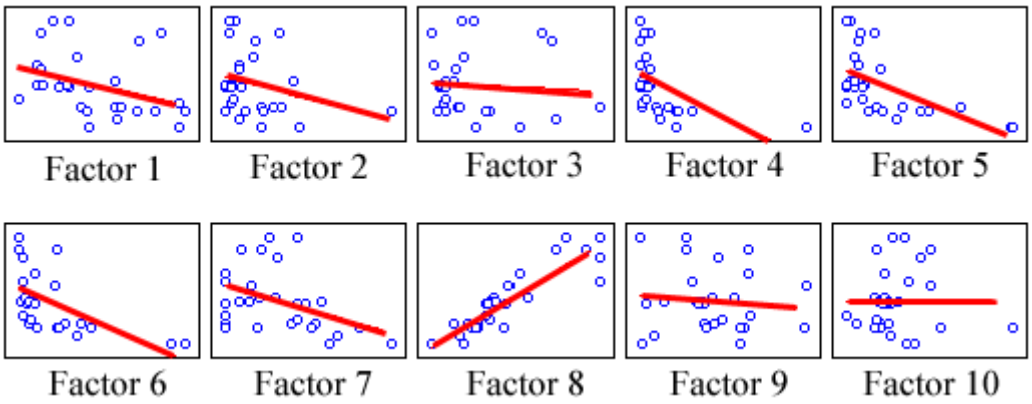

Figure 1. Scatterplots of scores versus factors

The scatterplots in Figure 1 of this response include the schools contributing to the summary analysis of the article, in its Figures 2-8 through 2-10. The ordinate of each scatterplot is the school-averaged test score for 1999. Although several factors have significant correlations, Factor 8 provided dominant and robust association of variance in a multifactor model. In the article, this factor is identified as "Per-capita community income (1989)." Another factor proportional to community income would certainly act as an effective substitute.

Willson and Kellow maintain that using "aggregate measures" may yield "misleading conclusions." Possibly—but from what kind of data are the test-based school rating systems themselves constructed? "Aggregate measures," I believe. Willson and Kellow don't seem to be disturbed that all test-based school rating systems, those they like as well as those they don't, are subject to similar potentials for distortion.

In their objections about "mixing levels of analysis," Willson and Kellow do not seem to have followed the article closely. Preliminary analyses through Table 2-8 use only school-based data. Intermediate steps, introducing per-capita incomes from the census, warn about problems from mixing levels. The summary analysis, Section 2.D, presents in Tables 2-20 and 2-21 and in Figures 2-8 through 2-10 the results from only communities with a single high school. For those schools, there is only one level. The summary analysis has no level mixing.

Despite sensitivities over "mixing levels of analysis," Willson and Kellow sometimes seem to confound issues of evaluating personal test scores in the context of personal factors with issues of evaluating test-based school ratings in the context of school and community factors. The article is focused on the latter topic, not the former.

Willson and Kellow say that measures such as percentages of students qualified for free or reduced-price lunch would be "more appropriate" than per-capita income for the purposes of the article—an opinion they don't explain or defend. The article shows that test scores are closely associated with incomes at all income levels found in the communities studied. Its results suggest that community income, as distinct from family income, may have substantial influence. A study of individuals would be needed to resolve the influences.

When typical incomes were well beyond poverty levels, as in many communities studied for the article, the percentage of high-school students qualified for free or reduced-price lunch became a poor proxy for income. A model using income directly was statistically much more effective. Willson and Kellow argue that this would be less of a problem for elementary-school students, but they neither cite nor present evidence. In a study focused on lower-income Florida communities Tschinkel, 1999, did find an association between test scores and "supported lunch" that was about as strong as the article finds between test scores and income for Massachusetts communities near Boston.

Willson and Kellow review MCAS tests and AREA/APS *Standards* (Committee, 1985, 1999), but the article does not invoke those or any other institutional standards. It is not focused on testing at a personal level, as standards are. Instead it is concerned with school ratings based on test scores. However, in this context Willson and Kellow are surely aware that jurisdictions such as California and Chicago ignored some of the AREA/APS standards in using commercial achievement tests for promotion and graduation tests. Other jurisdictions such as Texas and Massachusetts claim they comply with those standards but have legalistic interpretations.

Scattterplot data from Texas that Willson and Kellow show in their first figure could not be compared with data from Massachusetts in the article, because Willson and Kellow did not provide data, cite a source from which to obtain data, or translate their "economic disadvantage" index to income. They do not say whether their index reflects household or community income, nor do they evaluate the accuracy of the index as an

income proxy for communities with typical incomes well above poverty.

There is a further, critical problem in trying to compare results from the Texas "accountability" testing program with those from Massachusetts. As many experiences show, teachers, students and parents adapt defensively to testing programs—the higher the stakes, in general, the stronger the defenses. Education agencies also respond in a variety of ways to public reactions. Massachusetts patterns from 1998-1999 might be compared with Texas patterns from 1985-1986, when Texas testing began, but Massachusetts tests from 1998-1999 more closely resemble contemporaneous Texas tests than the Texas tests from 1985-1986.

Texas started "accountability" testing in 1985 and is now more than ten years into a second generation of tests (TAAS). It has been enforcing a graduation test requirement and maintaining its Accountability Rating System for more than eight years. By the late 1990s, there were widespread reports of weeks spent on test cramming, of "TAAS rallies," of heavy school spending on test prep materials and consultants, and of scandals over falsifying reports (see McNeil, 2000, for examples). Some observers such as Haney, 2000, suspect dropout rates have been redefined to conceal problems.

Massachusetts started "accountability" testing in 1998. Its graduation requirement will not directly affect any student until 2003. Massachusetts has some communities with a history of aggressive testing, such as Worcester; but before the school year ending in 2000 most schools made light to moderate responses to MCAS tests, and the Board of Education discounted most concerns about higher dropout rates.

Baseline data for "Significance of Test-based Ratings" were from 1998 and 1999, the quietest years of the MCAS program. Differences between those years were used to estimate variability, and the 1999 scores were used for most of the effects models. These data reflect conditions of Massachusetts schools before most strong responses. Scores from 2001, unavailable when the study was conducted, clearly show effects of strong responses, which will probably grow. Under state threats to take over or close low-scoring schools, there have already been heavy efforts to increase scores in some schools, involving test cramming that would be familiar to Texans; and there have been widely reported score increases.

In a short paragraph after their Texas scatterplot data, Willson and Kellow again object to correlating income with test scores, claiming income to be "uninterpretable." One is reminded of Tevye from *Fiddler:* "Impossible! Impossible!" Of course, the interpretation is entirely possible. Grissmer, et al., 2000, do it, the Nader organization does it, the Educational Testing Service does it, and if they try a bit harder, Willson and Kellow can do it too.

Willson and Kellow advocate teaching and learning metrics sometimes called the "value-added model" (e.g., by McLean, et al., 1998). Sanders has popularized a variant of this approach, and he supports it commercially (Sanders & Horn, 1995). The stability and significance of ratings based on such methods have recently been questioned by Kane & Staiger, 2001b, who also estimate contributions to score volatility from several sources.

Consistent with the article's observation (not "error") under Figure 2-10, Willson and Kellow also find year-to-year score changes exhibiting low significance. The

Massachusetts Department of Education uses a score-change metric (Mass. DoE, 1999) slightly more robust than year-to-year changes. Kane & Staiger, 2001a, propose filters applied to several years of scores for better metrics. Willson and Kellow might compare score volatility estimates, by sources, that were obtained by Kane & Staiger, 2001b, from North Carolina elementary school testing against the within-year and between-year score variations from Texas testing.

Willson and Kellow complain that the article does not explore the content of MCAS tests. As previously stated, the article's focus is on significance of test scores for constructing school ratings, not on internal properties of the tests. However, the article provided Willson and Kellow with references to the full MCAS test content (Mass. DoE, 2000a, for 2000) and technical manual (Mass. DoE, 2000b, for 1999), which Massachusetts publishes on the Internet. They have had unlimited access to these documents for any reviews they find "appropriate."

In their Table 3, Willson and Kellow present a correlation matrix of score changes from Texas elementary schools plus social and school factors. Again it was not possible to compare Texas data with Massachusetts data from the article. The article associates factors with scores, while Willson and Kellow associate them with score changes for cohorts of identified students. There are also at least two major problems with results that Willson and Kellow present. First, they have score changes with high volatility. Obtaining a robust pattern would require multiple years, perhaps with filters such as Kane & Staiger, 2001a, propose. Second, they have social and school factors with substantial correlations, but they offer no multifactor model and no stepwise or combinatorial analysis of variance.

Willson's and Kellow's title for their critique echoes a shopworn slogan of the education hustlers: that school-based standard test scores are sending us a "message." Other than a bundle of sticks, what might that message be? A question rarely asked about "accountability" programs is whether their tests measure anything useful. What they measure is whether a test-taker, in a constrained situation, can interpret isolated fragments of information, solve small, arbitrary puzzles, recall miscellaneous items, or write in a simplistic style. Otherwise such situations may rarely be encountered—except perhaps with crossword puzzles or quiz shows.

Life's challenges are hardly ever so neatly packaged as the questions on a school-based standard test. They are often far more difficult: grasping people's real wants and needs, seeing advantages where others see limitations, organizing experience to make sense of it, understanding one's own blind spots, persisting against adversity, motivating people and guiding them. In most circumstances other than getting certificates that depend on results from those tests, how would the results be of practical use? Attempts to measure education effectiveness using the current generations of state "accountability" tests may be mansions built on sand.

Associations of incomes with "aptitude" test scores have been recognized in the U.S. for more than 80 years. There are related studies about effects of poverty on cognitive development, such as Smith, et al., 1997, but the underlying behaviors at higher incomes are not understood much better now than they were in the 1920s. Flynn, 1984, showed that average IQ scores have been rising dramatically over time, suggesting that the underlying behaviors involve training or experience. Recently Dickens & Flynn, 2001, proposed an interpretive model, but so far little investigation of it has been reported.

State "accountability" tests may be heavily weighted for the same cognitive skills that are sampled by "aptitude" tests, leading to associations with income like those that "Significance of Test-based Ratings" finds.

As this article shows, Massachusetts school ratings, based solely on MCAS test scores, are not precise educational measures of high significance. Score variations are large, and scores appear to reflect heavily the social advantages or disadvantages that students bring to schools from their backgrounds, not necessarily the effectiveness of the schools themselves.

References

Bolon, C. (2001). Significance of test-based ratings for metropolitan Boston schools. *Education Policy Analysis Archives 9*(42). October 16, 2001, available at http://epaa.asu.edu/epaa/v9n42.

Committee to Develop Standards for Educational and Psychological Testing, Novick, M. R., Chair (1985, revised 1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association and American Psychological Association.

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: the IQ paradox resolved. *Psychological Review 108*(2), pp. 346-369.

Flynn, J. R. (1984). The mean IQ of Americans: massive gains 1932 to 1978. *Psychological Bulletin 95,* pp. 29-51.

Grissmer, D. W., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving Student Achievement: What NAEP State Test Scores Tell Us.* Santa Monica, CA: RAND Corp.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives 8*(41) at http://epaa.asu.edu/epaa/v8n41.

Kane, T. J., & Staiger, D. O. (2001a). *Improving School Accountability Measures.* Cambridge, MA: National Bureau of Economic Research, Working Paper 8156.

Kane, T. J., & Staiger, D. O. (2001b). *Volatility in School Test Scores.* Washington, DC: Brookings Institution.

Massachusetts Department of Education (1999). *School and District Accountability System*.

Massachusetts Department of Education (2000a). *Release of Spring 2000 Test Items.* Available at http://www.doe.mass.edu/mcas/2000/release/.

Massachusetts Department of Education (2000b). *MCAS 1999 Technical Report.* Available at http://www.doe.mass.edu/mcas/2000/news/pdf/99techrep.pdf.

McLean, J. E., Snyder, S. W., & Lawrence, F. R. (1998). A school accountability model. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, New Orleans, Louisiana, November 4-6, 1998.

McNeil, L. M. (2000). *Contradictions of School Reform,* New York, NY: Routledge.

Milton, S (1999, February 25). Town lawyers endorse funding of RAGE lawsuit. *Cape Cod Times,* Barnstable, MA.

Sanders, W. L., & Horn, S. P. (1995). Educational assessment reassessed. *Education Policy Analysis Archives 3*(6) at http://epaa.asu.edu/epaa/v3n6.html.

Smith, J. R., Brooks-Gunn, J., & Klebanov, P. K. (1997). Consequences of living in poverty for young children's cognitive and verbal ability and early school development. In G. J. Duncan & J. Brooks-Gunn (Eds.), *Consequences of Growing Up Poor* (pp. 132-167). New York, NY: Russell Sage Foundation.

Texas education funding challenged in new lawsuit (1998, May 8). *Daily Tax Report.* Washington, DC: Bureau of National Affairs.

Tschinkel, W. R. (1999, April 25). Poverty, not bad schools, hinders learning. *Miami Herald.*

About the Author

Craig Bolon
Planwright Systems Corp
P.O. Box 370
Brookline, MA
02446-0003 USA
Email: cbolon@planwright.com
Phone: 617-277-4197

Craig Bolon is President of Planwright Systems Corp., a software development firm located in Brookline, Massachusetts, USA. After several years in high energy physics research and then in biomedical instrument development at M.I.T., he has been an industrial software developer for the past twenty years. He is author of the textbook *Mastering C* (Sybex, 1986) and of several technical publications and patents. He is an elected Town Meeting Member and has served as member and Chair of the Finance Committee in Brookline, Massachusetts.

EPAA Editorial Board

EPAA Spanish Language Editorial Board